“Big Data” Projects in High Energy Physics and Cosmology at Kansas State University

Glenn Horton-Smith, Associate Professor, Department of Physics,
Kansas State University

Motivations, necessities, and methods of “big data” analysis in High Energy Physics

The goal of high energy physics (HEP) research is to discover as much as possible about the elementary properties of energy, matter, space, and time. New discoveries are made by analyzing data from new experiments performed under conditions allowing the observation of phenomena that could not be seen in previous experiments. Present-day experimental high energy physics has been characterized as having three frontiers[i]: an Energy Frontier, explored by experiments requiring the highest energies achievable; an Intensity Frontier, explored by experiments requiring the highest intensities achievable; and a Cosmic Frontier, explored using naturally-occurring cosmic particles and observations of the cosmos. As will be explained, research at these frontiers naturally requires the analysis of vast amounts of data. The HEP research program at Kansas State University (K-State) will be used an example.

The HEP group at K-State engages in research on all three frontiers. On the Energy Frontier, the primary effort is the CMS experiment[ii] at the Large Hadron Collider (LHC), whose goals include study of the Higgs boson and discovery of new particles and other phenomena. On the Intensity Frontier, we work on multiple neutrino experiments[iii], whose goals include the understanding of the nature of mass and the study of matter/antimatter asymmetries. On the Cosmic Frontier, the emphasis is on developing and testing models of dark energy[iv], and alternatives thereto, with the goal of understanding the nature of the phenomenon driving the observed acceleration of the expansion of the universe.

The CMS experiment requires the high energy particle collisions of the LHC to produce Higgs bosons and to test other hypotheses such as super-symmetry and extra dimensions. Only a small fraction of the collisions produce phenomena of interest. The raw data is therefore dominated by signals from known phenomena already explored at lower energies.

In real time, there are 20 million collisions per second producing signals in a detector with many millions of raw signal channels. Permanently storing data from scores of trillions of digitized signals every second is infeasible. Instead, the data from CMS is reduced in multiple stages by using “triggers” in real time to reduce the recorded data to the order of 10 petabytes per year (1 petabyte = 10^{12} bytes). Later data-reduction stages are applied to the recorded data to identify the particle tracks seen in each event and produce smaller data sets that are richer in interesting events. The data is stored and processed on the CMS Computing Grid[v], which is organized into "tiers", with lower-numbered tiers storing and analyzing the less-processed data, and higher-numbered tiers working on output from the lower-numbered tiers.

In contrast, neutrino experiments require high intensities because neutrinos have extremely low interaction probabilities. (To give an often used illustration, if the sun could be surrounded by a light year of solid lead, a large fraction of the neutrinos produced in the sun would still escape.) The hardware-level trigger rate varies greatly between neutrino experiments, but is invariably much lower than collider experiments, and typically on the order of 1 to 1000 triggers per second, dominated by non-neutrino sources of "background" events. Neutrino rates are typically in the range of 10^{-5} to 10^{-3} per second, or one to a hundred per day. Like the collider experiments, neutrino experiments search for relatively rare events in a much larger data set.

The number of channels in neutrino experiments tends to be of the order of thousands or tens of thousands, much smaller than in collider experiments. That fact, along with the lower total trigger rate, allows collecting all data to disk in real time, with all analysis done later. An experiment such as KamLAND or Double Chooz might write on the order of 0.1 petabyte/year.

On the Cosmic Frontier, the phenomena investigated are too weakly interacting, too rare, or too energetic to be studied using artificial sources. The kinds of observations analyzed for Cosmic Frontier research include multiple high resolution images searched for distant objects (e.g., distant galaxies) and particular types of time variations (e.g, supernovae or gravitational lensing). The data sets here are large because the universe is so big, and time-varying phenomena so transient: lots of images with many pixels are needed. The scientists who build and operate astronomical instruments perform basic analyses that are published as results of large astronomical "surveys". The K-State cosmology group under Prof. Bharat Ratra primarily concentrates on theory, and analyzes what the astronomical survey results mean to theoretical models.

A common feature in all the research described above is that we obtain information, with quantified uncertainties, from large data sets that have been subjected to strict selection criteria. Necessary analytic skills include:

- Reconstruction/identification: transforming raw data into "physics objects."
- Simulation/modeling: obtaining simulated data as it would be for a given model.
- Evaluation of uncertainties, significance, coverage regions for these experiments.

Some tools and methods of "big data" analysis in High Energy Physics

In HEP, we tend to use open-source software as much as possible. The ability to inspect source code, and correct and contribute to it if necessary, is important. Two examples of commonly used software are Geant4[vi] and ROOT[vii].

Geant4 is a standard software library for creating models of particle detectors. The primary purpose of such models is to correctly calculate the interactions of particles passing through the detector and the detectable signals (*e.g.*, ionization or light) produced by those interactions. The visualization of detector geometry is provided as a tool for debugging the implementation of detector geometry; an example is shown in Fig. 1.

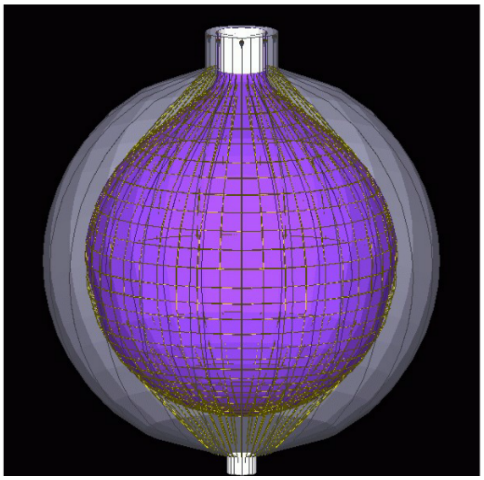

Figure 1: Part of the KamLAND detector model in Geant4 [vi]

The ROOT object-oriented data analysis framework is perhaps the most common tool for data analysis and visualization in HEP. It provides features similar to other data analysis packages, including functions and objects for storing and retrieving data sets, generating graphs, plots, and histograms, generating random numbers and distributions, fitting the data, and various means for implementing custom analyses in C++ or other programming languages. An example of a fitted histogram made in ROOT is shown in Fig. 2.

The way in which the programmability feature is implemented sets ROOT apart from many other data analysis software tools. ROOT is both an interactive tool and a software library that can be used in any C++ program. The interactive capabilities include a graphical user interface, but ROOT also has a command-line interface that can access every function in the library, using C++ syntax. Like many tools, ROOT has a scripting feature, but ROOT's scripting language also uses C++ syntax. This allows a process of analysis development leading from small to big data analysis that can proceed as follows:

1) Try something interactively in

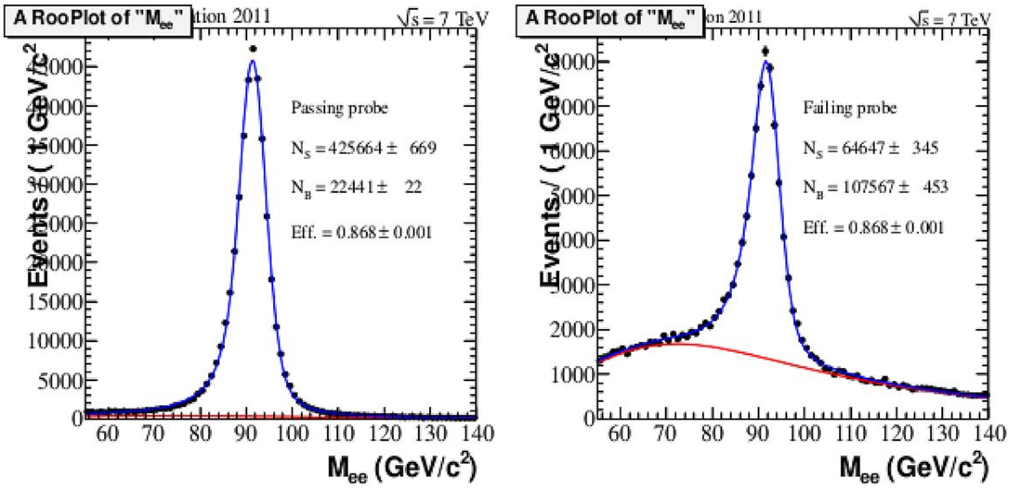

Figure 2: An example of fits performed to histograms as part of a tag-and-probe analysis, from [viii]. Fits and plots were done using ROOT software.

ROOT.

2) Copy the interactive commands into a ROOT "script" and run it interactively.
3) Rewrite the script in the form of a proper C++ function. Load it interactively and run the function from ROOT.
4) Rewrite it so it is a complete, compilable C++ file. Compile and load from ROOT, run the function. (At this point, one has natively compiled code that runs quickly and can be run on nodes in a compute farm.) One can also compile the same file outside of ROOT and use it in any C++ program.

Intermixed with this development process is a process of presentation of ideas and intermediate results to individual colleagues and groups of various sizes within the experimental collaboration, invariably leading to suggestions and corrections based on the colleagues' knowledge of relevant aspects of the experiment. The design of ROOT allows the researcher to quickly modify and repeat analyses as needed.

A great number of analyses, with associated plots and histograms, are used to validate models and present work to collaborators and the world. Each analysis has unique aspects. Two particularly important aspects of HEP are obtaining reliable measurements of data selection efficiency and estimating backgrounds. In this context, selection efficiency is the fraction of events of a desired type that survive the triggers and selection cuts, and backgrounds are any events of undesired type that remain in the sample after the selection cuts. Data-driven methods are preferable to simulation in making efficiency measurements and background estimates. Monte Carlo (MC) simulations based on modeling of the detector and the physics under study can be useful, but the reliability of the MC must be established using data-driven methods.

A particularly useful data-driven method for measuring efficiency is the “tag-and-probe” method. It is especially useful when the new particles or interactions are detected solely through the observation of known particles whose properties are well understood. The known particles are also produced in simpler, well-understood reactions. The tag-and-probe method “tags” known interactions in which a particle of a particular type must be produced, then uses the particle known to be produced in that interaction as a “probe” to determine efficiency and an estimated uncertainty for the efficiency estimate.

In order to eliminate false signals from the “tag” while not biasing the “probe”, it is important to choose a “tag” interaction that can be selected with very tight criteria overall but loose criteria on the probe particle. Often this can be done by using interactions that produce particles of a given type in pairs, and applying tight selection cuts to only one particle in the pair.

To determine the efficiency of selection for a hypothesized new particle, the tag-and-probe analysis is performed for each type of particle that would appear in the decay of the new particle. A nice example of such an analysis can be found in the dissertation of Irakli Svintradze[viii], which happened to be the most recent K-State HEP dissertation to be completed before this workshop. Two histograms used in evaluating one efficiency factor in the dissertation are shown in Fig. 2. Generally speaking, the product of the efficiencies does not directly represent the efficiency of selection for decays of the new particle, so the analysis is performed both on data from the detector and on MC simulations of tag interactions. If the MC is reasonably accurate, the efficiencies from data and MC will be very close. Any slight differences can be applied as corrections to the efficiencies of MC simulations of the new particle decay, thus obtaining an efficiency estimate based on the modeled properties of the hypothesized particle and reliable, data-driven estimates for the detection efficiencies of every secondary particle.

There are many other issues besides selection efficiency that HEP experimentalists consider when analyzing big datasets, two of which are the so-called "look elsewhere" effect when searching over a wide region of some parameter (*e.g.*, energy) for a signal instead of at a single predetermined value, and the effect of tails of statistical distributions that might cause an uninteresting but prevalent phenomenon to look like something interesting but rare. A variety of techniques have been developed to address these issues in an unbiased way, one example of which is the closed-box (or "blind box") analysis in which one or more parameters of signal-like events are hidden from use in any analysis until after all steps of the analysis are completed except the final determination of the signal or parameter of interest.

It is not hard to think of other contexts in which these issues are important. Statisticians and analysts from other fields are well aware of these issues in general. However, HEP experimentalists have been dealing with huge data sets for a long time, and consideration of HEP practices and techniques may provide unique perspectives and ideas.

Further thoughts

The previous two sections cover the content presented at the 2013 Merrill Workshop. Here are a few notes on points touched on in discussions afterwards regarding similarities and differences between HEP analysis techniques and data analysis in other contexts.

On tags and probes: In the context of evaluating scholarly output, Hirsch's original paper proposing the *h*-index[ix] used the Nobel Prize in Physics (and other awards) as a kind of "tag" and the winning physicists as "probes" to suggest a threshold for comparable levels of scientific impact and relevance. The study of the *h*-index using prize-winning scientists is significantly less sophisticated than tag-and-probe analysis as used in high energy physics analyses due to the lack of a model for what actually produces a Nobel-caliber physicist and the relatively small number of quantities used in determining the *h*-index. To be fair, Hirsch only proposed the *h*-index as "a useful index".

On the use of pairs of identical objects: Studies of twins are useful in the social and medical sciences. However, it is difficult to do this in analyzing academic performance data such as the *h*-index, lacking a sure way of producing pairs of researchers of equal impact and relevance.

On closed-box analysis: In academic analytics, data from "peer" and "aspirational peer" institutions and programs can be used to enable a kind of closedboxed analysis in which metrics are developed in a data-driven way without using any data from the analyst's own institution. Insisting on such an approach to academic analysis could be a way for top research administrators to address concerns about releasing detailed program data to individual program heads or researchers for their own analyses.

I thank the Merrill Foundation for supporting this most interesting workshop, the organizers for the excellent way it was run, and all the participants for the great discussions.

References

i. Particle Physics Project Prioritization Panel (P5-2008), C. Baltay, Chair, US Particle Physics: Scientific opportunities: A strategic plan for the next ten years, 2008. Available at http://science.energy.gov/~/media/hep/pdf/files/pdfs/p5_report_06022008.pdf, browsed 2013/08/20.

ii. The CMS Collaboration et al, The CMS experiment at the CERN LHC, JINST 3 S08004, 2008. doi:10.1088/1748-0221/3/08/S08004

iii. Double Chooz: Y. Abe et al. [Double Chooz collaboration], Phys.Rev.Lett., 108:131801, 2012; ArgoNeuT: C. Anderson, et al. [ArgoNeuT collaboration], JINST 7, P10019 (2012); MicroBooNE: The MicroBooNE collaboration, et al., MicroBooNE-doc-1821-v13 (2012); LBNE: The LBNE Conceptual Design Report (August 2012), available from http://lbne.fnal.gov/about_LBNE.shtml.

iv. P.J.E. Peebles and B. Ratra, The Cosmological constant and dark energy, Rev.Mod.Phys. 75 (2003) 559-606. doi:10.1103/RevModPhys.75.559

v. The CMS Collaboration et al, CMS: The computing project technical design report, CERN report CERN-LHCC-2005-023, 20 June 2005. Available at http://cds.cern.ch/record/838359/files/lhcc-2005-023.pdf, browsed 2013/08/20.

vi. The GEANT4 collaboration, et al, NIM A 506 (2003) 250-303, and IEEE Trans.Nucl.Sci. 53 No. 1 (2006) 270-278. See also http://geant4.cern.ch.

vii. Antcheva, et al., ROOT: A C++ framework for petabyte data storage, statistical analysis and visualization, Comput.Phys.Commun. 180 (2009) 2499-2512; also available as FERMILAB-PUB-09-661-CD at http://lss.fnal.gov/cgi-bin/find_paper.pl?pub-09-661. See also http://root.cern.ch.

viii. Svintradze, Diboson Physics with CMS Detector, PhD dissertation, Kansas State University, 2013. Available at http://krex.k-state.edu/dspace/handle/2097/15782.

ix. J.E. Hirsch, An index to quantify an individual's scientific research output, Proc.Nat.Acad.Sci. 46 (2005) 16569. doi:10.1073/pnas.0507655102